LECTURE 13. FURTHER APPLICATIONS OF THE CANONICAL FORMALISM (CONT.'D)

• <u>Cooperative Binding</u>

The general model of ligand-receptor binding that we introduced in the previous lecture was found to provide a highly satisfactory description of the binding of oxygen to myoglobin. But that model would fail to adequately describe the binding of oxygen to haemoglobin. Haemoglobin has 4 binding sites for oxygen, and in experimental studies of the binding of increasing amounts of oxygen to a fixed concentration of haemoglobin, the graph of fractional occupancy versus oxygen partial pressure has the following qualitative appearance

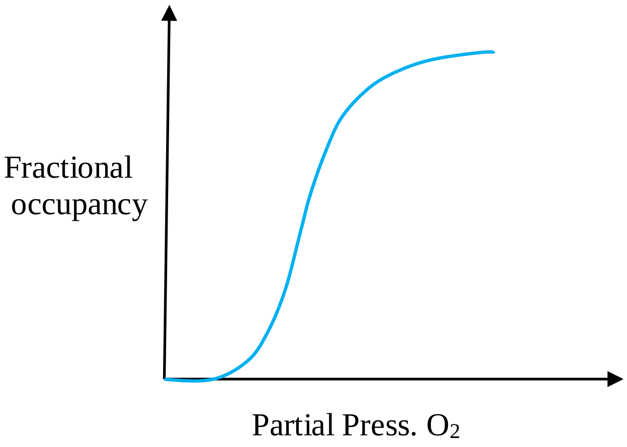


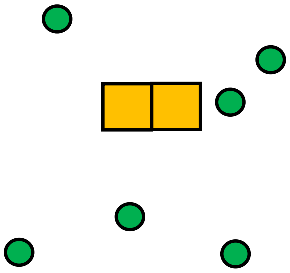
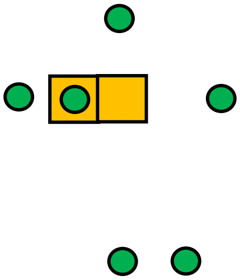
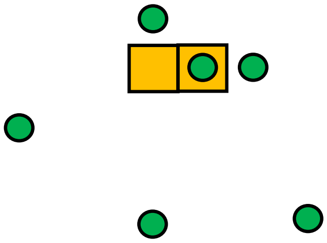
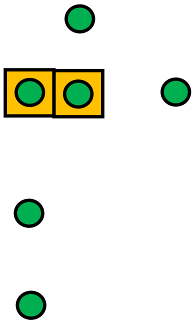
Fig. 1

which is quite different from the binding curve we had studied earlier. The S-shaped nature of the curve in Fig. 1 is usually ascribed to "cooperativity", a phenomenon in which the binding of the first O_2 molecule to a binding site increases the likelihood of the binding of subsequent O_2 molecules to the remaining sites. The molecular basis for cooperativity appears to be a conformational change induced by the binding of the first O_2 molecule; this leads to an altered binding energy for the subsequent binding events.

Can we find a statistical mechanical rationale for cooperative binding?

To address this question, let's explore what happens if we make a few minor modifications to our earlier model of ligand-receptor binding. In that model, we imagined that L ligands could move freely between N lattice sites, one of which was occupied by a receptor. We assumed that the energy of a ligand was ε_s when it occupied an empty lattice site and that it was ε_b when it occupied the receptor site. Now let's introduce some other assumptions: (i) that the receptor has 2 binding sites (haemoglobin has four, but we'll consider just two to keep the math simple), (ii) that when either of these sites is singly occupied, the energy is ε_b, but that when both are occupied (simultaneously), the energy is further changed by an amount J.

With the model defined, our strategy for finding the probability that both sites on the receptor are bound is the same: identify the possible states of the system (its macrostates, specifically), determine their energies, and calculate their multiplicities. The results of this exercise are shown in the table below: (the lattice cells have been omitted.)

State	Energy	Multiplicity
1. 	$L\varepsilon_s$	$\dfrac{N!}{L!(N-L)!} \approx \dfrac{N^L}{L!}$
2. 	$(L-1)\varepsilon_s + \varepsilon_b$	$\dfrac{N!}{(L-1)!(N-L+1)!} \approx \dfrac{N^{L-1}}{(L-1)!}$
3. 	$(L-1)\varepsilon_s + \varepsilon_b$	$\dfrac{N!}{(L-1)!(N-L+1)!} \approx \dfrac{N^{L-1}}{(L-1)!}$
4. 	$(L-2)\varepsilon_s + 2\varepsilon_b + J$	$\dfrac{N!}{(L-2)!(N-L+2)!} \approx \dfrac{N^{L-2}}{(L-2)!}$

The probability p_2 that both receptor sites are occupied is now easily calculated from our general statistical mechanical formalism (cf. Eq. (5)):

$$p_2 = \frac{\dfrac{N^{L-2}}{(L-2)!} e^{-\beta[(L-2)\varepsilon_s + 2\varepsilon_b + J]}}{\dfrac{N^L}{L!} e^{-\beta L \varepsilon_s} + 2\dfrac{N^{L-1}}{(L-1)!} e^{-\beta[(L-1)\varepsilon_s + \varepsilon_b]} + \dfrac{N^{L-2}}{(L-2)!} e^{-\beta[(L-2)\varepsilon_s + 2\varepsilon_b + J]}}$$

Pulling out a factor of $N^L e^{-\beta L \varepsilon_s} / L!$ from the denominator, p_2 simplifies to

$$p_2 = \frac{\dfrac{L(L-1)}{N^2} e^{-2\beta\Delta\varepsilon - \beta J}}{1 + 2\dfrac{L}{N} e^{-\beta\Delta\varepsilon} + \dfrac{L(L-1)}{N^2} e^{-2\beta\Delta\varepsilon - \beta J}}$$

$$\approx \frac{\left(\dfrac{c}{c_0}\right)^2 e^{-2\beta\Delta\varepsilon - \beta J}}{1 + 2\dfrac{c}{c_0} e^{-\beta\Delta\varepsilon} + \left(\dfrac{c}{c_0}\right)^2 e^{-2\beta\Delta\varepsilon - \beta J}}$$

where as before $\Delta\varepsilon = \varepsilon_b - \varepsilon_s$. This is our sought-for expression. It contains essentially two adjustable parameters, $\Delta\varepsilon$ and J, and if we were to assign the following values to them, $\Delta\varepsilon = -5 k_B T$ and $J = -0.25 k_B T$, and then plot p_2 versus c/c_0, our graph would qualitatively reproduce the S-shaped curve of Fig. 1.

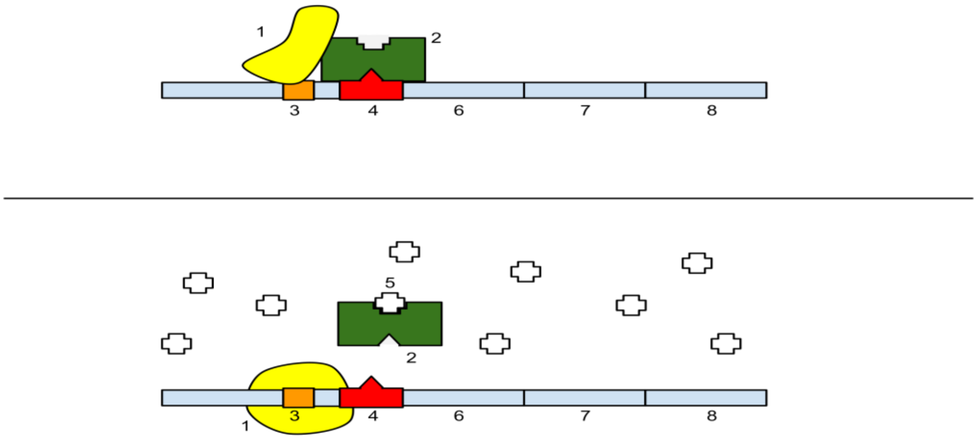
• The Regulation of Gene Expression

The ligand-receptor binding model of the previous sections is actually quite versatile, and it can be applied to other biological processes as well. One of these processes is gene expression, which refers broadly to the sequence of steps by which information in DNA is copied onto RNA and then translated into protein. The process is highly regulated, meaning there are a number of chemical reactions along its pathway that can turn it on or off. The binding of an RNA polymerase molecule to a specific site on the DNA (called a promoter) is one such reaction – it turns gene expression on. The binding of another kind of protein molecule – a repressor – to the same site is another such reaction – it turns the process off. Under conditions where significant numbers of polymerase and repressor molecules move randomly around the DNA, the competition between these two reactions and reactions where polymerase and repressor can also bind to other sites on the DNA that have no effect on gene expression, is what controls how far this process is carried forward, and how much protein is generated thereby.

A schematic of the regulatory mechanism in *E. Coli* that controls the metabolism of the sugar lactose is shown in the two panels below. Lactose is a fuel source that's hydrolyzed to glucose and galactose by the enzyme β-galactosidase, which is one of three enzymes encoded by three *E. Coli* genes (the grey regions numbered 6, 7 and 8 in the figure.) Since it's generally wasteful for the bacterium to produce enzymes when there's no lactose around, a repressor (the green shape, numbered 2) occupies a site on the DNA (the red region, numbered 4) that prevents RNA polymerase (the yellow blob, numbered 1) from fully occupying a promoter site (the orange rectangle, numbered 3) that when activated would have led to the initiation of transcription.

When lactose molecules are abundant (the white shapes numbered 5 in the second panel), they bind to the repressor, which then unbinds from the DNA, allowing the polymerase to now fully occupy the promoter site, thereby initiating gene expression and then eventually producing the enzymes that digest the lactose.

A neat bit of molecular choreography!



A number of questions can now be asked. For instance, as a function of repressor concentration, what is the probability that the promoter site is occupied by RNA polymerase? There are experimental data on *E. coli* that bear on this question. They are typically expressed in terms of a quantity called the fold-change of gene expression, *f*, which is defined as the concentration of enzyme produced in the *presence* of repressor divided by the concentration of enzyme produced in the *absence* of repressor. Since the amount of enzyme produced depends on the likelihood of the promoter site being occupied by RNA polymerase, *f* can be written as

$$f = \frac{p_b(R \neq 0)}{p_b(R = 0)} \qquad (6)$$

where p_b is the binding probability of polymerase and R is the number of repressor molecules.

Can we use statistical mechanics to obtain an expression for f ? To address this question we first need a model of gene expression and regulation. One such model is described in the next section.

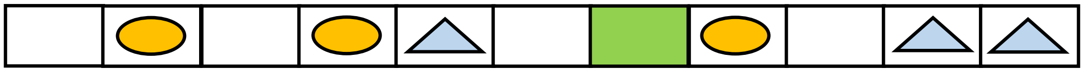
• <u>A Three-State Model of Gene Expression</u>

Any model of gene expression must include at least 3 components: DNA, polymerase and repressor. We can model DNA as a long linear ribbon having *M* binding sites (depicted as open cells in the figures below), one of which (the green cell) is a promoter site. *M* is expected to be on the order of 10^6, the typical number of base pairs in a bacterial genome. Molecules of RNA polymerase (*P* for short, and depicted as orange ovals) and repressor (*R*, blue triangles) can be modelled as particles that move freely across the DNA and occupy its cells at random. A snapshot of a section of DNA that has 3 *P* molecules and 3 *R* molecules bound to it at *non*-promoter sites is shown in the first panel. When *P* binds to the promoter site (i.e., when it occupies the green cell), a gene somewhere on the DNA is assumed to be turned on, and when *R* binds to it, the gene is assumed to be turned off. Suppose that the solution around the DNA molecule (at temperature *T*) contains N_P polymerases and N_R repressors, with both N_P and N_R on the order of 100. Let the binding energy between *P* and the promoter site be ε_P^s (*s* denoting specific) and that between *P* and a non-promoter site be ε_P^{ns} (*ns* denoting non-specific.) Let the corresponding binding energies for *R* be ε_R^s and ε_R^{ns}. Assume no multiple occupancy of binding sites.

Because there is a direct correlation between gene expression and the binding of *P* to the promoter, the goal now is to calculate how likely this binding event will be.

As we saw in the example of ligand-receptor binding, to calculate a binding probability using statistical mechanics, we need to identify the various states that are permitted by our model, find their energies, and determine their multiplicities. For the present model of gene expression, there are 3 energy macrostates. They correspond to the following situations: (i) All *P* and *R* molecules are bound to non-promoter sites, (ii) A single *P* is bound to the promoter site and all the remaining molecules to non-promoter sites, and (iii) A single *R* molecule is bound to the promoter site and all the remaining molecules to non-promoter sites. Typical realizations of these possibilities are shown schematically below for $N_P = N_R = 3$. The energies U and multiplicities Ω for general N_P and N_R are shown below the figures, along with the approximation for Ω applicable in the limit $M >> N_P, N_R$.

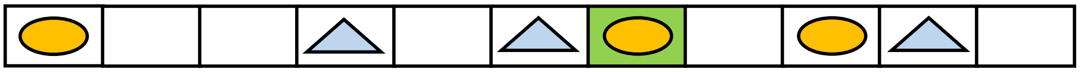
Macrostate 1



$$U_1 = N_P \varepsilon_P^{ns} + N_R \varepsilon_R^{ns}$$

$$\Omega_1 = \frac{M!}{N_P!N_R!(M - N_R - N_P)!} \approx \frac{M^{N_R+N_P}}{N_P!N_R!}$$

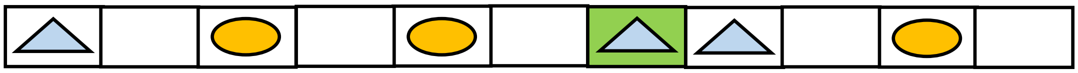
Macrostate 2



$$U_2 = (N_P - 1)\varepsilon_P^{ns} + \varepsilon_P^{s} + N_R \varepsilon_R^{ns}$$

$$\Omega_2 = \frac{M!}{(N_P - 1)!N_R!(M - N_R - (N_P - 1))!} \approx \frac{M^{N_R+N_P-1}}{(N_P - 1)!N_R!}$$

Macrostate 3



$$U_3 = N_P \varepsilon_P^{ns} + \varepsilon_R^{s} + (N_R - 1)\varepsilon_R^{ns}$$

$$\Omega_3 = \frac{M!}{N_P!(N_R - 1)!(M - (N_R - 1) - N_P)!} \approx \frac{M^{N_R-1+N_P}}{N_P!(N_R - 1)!}$$

From these expressions, it's now an easy matter to read off the probability that P binds to the promoter; recalling Eq. (5), it is given by

$$p_b = \frac{\dfrac{M^{N_R+N_P-1}}{(N_P-1)!N_R!} e^{-\beta[(N_P-1)\varepsilon_P^{ns}+\varepsilon_P^{s}+N_R\varepsilon_R^{ns}]}}{Q} \qquad (7)$$

where

$$Q = \frac{M^{N_R+N_P}}{N_P!N_R!} e^{-\beta[N_P\varepsilon_P^{ns}+N_R\varepsilon_R^{ns}]} + \frac{M^{N_R+N_P-1}}{(N_P-1)!N_R!} e^{-\beta[(N_P-1)\varepsilon_P^{ns}+\varepsilon_P^{s}+N_R\varepsilon_R^{ns}]} +$$

$$+ \frac{M^{N_R-1+N_P}}{N_P!(N_R-1)!} e^{-\beta[N_P\varepsilon_P^{ns}+\varepsilon_R^{s}+(N_R-1)\varepsilon_R^{ns}]} \qquad (8)$$

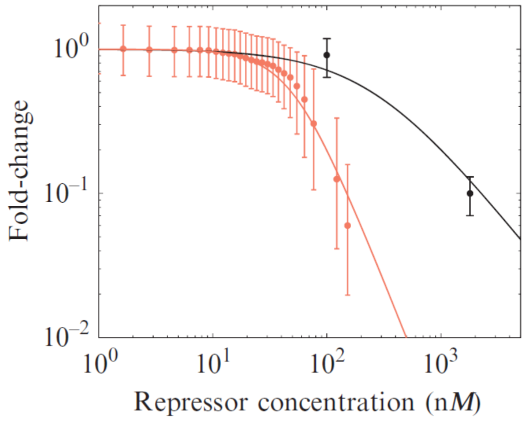
Pulling out a factor of $\frac{M^{N_R+N_P-1}}{(N_P-1)!N_R!} e^{-\beta[(N_P-1)\varepsilon_P^{ns}+\varepsilon_P^{s}+N_R\varepsilon_R^{ns}]}$ from the denominator, we can simplify Eq. (7) to

$$p_b(N_P,N_R,M) = \frac{1}{1+\frac{M}{N_P}e^{\beta(\varepsilon_P^s-\varepsilon_P^{ns})}\left[1+\frac{N_R}{M}e^{-\beta(\varepsilon_R^s-\varepsilon_R^{ns})}\right]} \qquad (9)$$

We're now in a position to calculate the fold change of gene expression, viz., $f = p_b(N_P,N_R,M)/p_b(N_P,N_R=0,M)$. From Eq. (9),

$$f = \frac{1+\frac{M}{N_P}e^{\beta(\varepsilon_P^s-\varepsilon_P^{ns})}}{1+\frac{M}{N_P}e^{\beta(\varepsilon_P^s-\varepsilon_P^{ns})}\left[1+\frac{N_R}{M}e^{-\beta(\varepsilon_R^s-\varepsilon_R^{ns})}\right]} \qquad (10)$$

If f is plotted versus N_R (suitably expressed as a concentration) for some chosen values of N_P/M, $\Delta\varepsilon_P \equiv \varepsilon_P^s - \varepsilon_P^{ns}$ and $\Delta\varepsilon_R \equiv \varepsilon_R^s - \varepsilon_R^{ns}$, and the results compared with data from studies of two different regulatory networks (Oehler et al. *EMBO J.* **13**, 3348 (1994) and Rosenfeld et al. *Science 307*, 1962 (2005)), the agreement is seen to be highly satisfactory (cf. the figure below.)



• <u>The Ideal Gas – Yet Again</u>

The ideal gas is a system that we have now almost beaten to death, but it’s worth re-examining with the canonical approach because this approach illustrates how much easier certain problems can be to solve when the right tools are deployed against them.

As before, we regard the ideal gas as a collection of *N* independent indistinguishable point particles confined to a box of volume *V*. Each particle can exist in one of a number of different quantum mechanical energy states defined by

$$\varepsilon_{l,m,n} = \frac{h^2}{8\mu L^2}(l^2 + m^2 + n^2), \quad l,m,n = 1,2,3,\ldots \tag{11}$$

In some microstate α of the entire collection of *N* particles, the energy U_α will therefore be

$$U_\alpha = \frac{h^2}{8\mu L^2}(l_1^2 + m_1^2 + n_1^2 + \cdots + l_N^2 + m_N^2 + n_N^2) \tag{12}$$

Assuming that the gas is in equilibrium with a thermal reservoir at temperature *T*, all we have to do now to determine the thermodynamic properties of the gas is to calculate *Q*, the canonical partition function, according to

$$Q = \frac{1}{N!}\sum_\alpha e^{-\beta U_\alpha} \tag{13}$$

As before, the reason for including a factor of *N*! in Eq. (13) is to account for the indistinguishability of the gas particles.

But how are we to interpret the “sum over states” $\sum_\alpha$? Since a microstate is a specification of 3*N* quantum numbers, different microstates are just different combinations of these quantum numbers, and all of these combinations will be realized if we write

$$Q = \frac{1}{N!}\sum_\alpha e^{-\beta U_\alpha} = \frac{1}{N!}\sum_{l_1=1}^{\infty}\sum_{m_1=1}^{\infty}\sum_{n_1=0}^{\infty}\cdots\sum_{l_N=1}^{\infty}\sum_{m_N=1}^{\infty}\sum_{n_N=1}^{\infty} e^{-\beta h^2(l_1^2+m_1^2+n_1^2+\cdots+l_N^2+m_N^2+n_N^2)/(8\mu L^2)} \tag{14}$$

Furthermore, since we’ve assumed that all the particles are independent of each other, the sums in Eq. (14) all factorize. That is,

$$Q = \frac{1}{N!}\sum_{l_1=1}^{\infty} e^{-\beta h^2 l_1^2/(8\mu L^2)} \sum_{m_1=1}^{\infty} e^{-\beta h^2 m_1^2/(8\mu L^2)} \sum_{n_1=0}^{\infty} e^{-\beta h^2 n_1^2/(8\mu L^2)} \times\cdots$$

$$\cdots\times\sum_{l_N=1}^{\infty}e^{-\beta h^2 l_N^2/(8\mu L^2)}\sum_{m_N=1}^{\infty}e^{-\beta h^2 m_N^2/(8\mu L^2)}\sum_{n_N=1}^{\infty}e^{-\beta h^2 n_N^2/(8\mu L^2)} \qquad (15)$$

Each of the separate sums in Eq. (15) is the same as every other, and so

$$Q=\frac{1}{N!}\left(\sum_{l_1=1}^{\infty}e^{-\beta h^2 l_1^2/(8\mu L^2)}\right)^{3N} \qquad (16)$$

All that remains now of the statistical mechanical part of the calculation is to perform a single summation. Unfortunately, this sum can't be done exactly, but it can be evaluated approximately without significant loss of precision when the system is of macroscopic dimensions. The approximation is to replace the sum by an integral:

$$Q=\frac{1}{N!}\left(\int_0^{\infty}dl_1 e^{-\beta h^2 l_1^2/(8\mu L^2)}\right)^{3N} \qquad (17)$$

Why this is a good approximation is because the successive terms in the summation in Eq. (16) differ so little from each other that the terms vary essentially continuously, and so for all practical purposes the sum is an integral. To see that the argument of the exponential in this equation hardly changes in going from l_1 to l_1+1, consider the case of an atom at room temperature with a mass μ of 10^{-22} g in a box of side $L=10\,\text{cm}$. For this system

$$\frac{\beta h^2(l_1+1)^2}{8\mu L^2}-\frac{\beta h^2 l_1^2}{8\mu L^2}=\frac{\beta h^2(2l_1+1)}{8\mu L^2}\approx(2l_1+1)\times10^{-20}\equiv\Delta$$

We've also seen that under standard temperature and pressure conditions, the quantum number l_1 is on the order of 10^9, so Δ is indeed extremely small, and no appreciable error results from replacing the sum by an integral.

The integral in Eq. (17) is well known $\left[\int_0^{\infty}dx e^{-ax^2}=\sqrt{\pi/a}/2\right]$, and after evaluating it, we're left with this result

$$Q=\frac{1}{N!}\left(\frac{2\pi\mu k_B T}{h^2}\right)^{3N/2}V^N \qquad (18)$$

The Helmholtz potential of the gas is therefore given by

$$F=-k_B T\left[-\ln N!+\frac{3N}{2}\ln\left(\frac{2\pi\mu k_B T}{h^2}\right)+N\ln V\right]. \qquad (19)$$

From the differential relation $dF = -SdT - PdV + \mu dN$, we can get the pressure as

$$P = -\left(\frac{\partial F}{\partial V}\right)_{T,N} \tag{20}$$

Differentiating Eq. (19) with respect to *V* at constant *T* and *N*, we see that $(\partial F / \partial V)_{T,N} = -Nk_B T / V$, and so

$$PV = Nk_B T \tag{21}$$

which is, of course, the ideal gas law.

It has obviously been much easier to derive this law within the canonical formalism than it was within the microcanonical formalism.